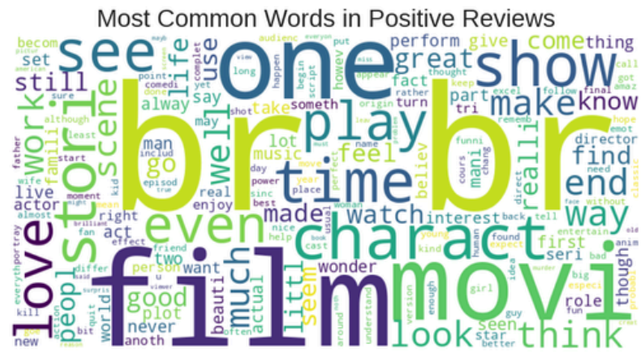
Machine Learning based classification for Sentimental analysis of IMDb reviews

Prateek P(PES1UG23AM211) | Vivian P(PES1UG23AM907)

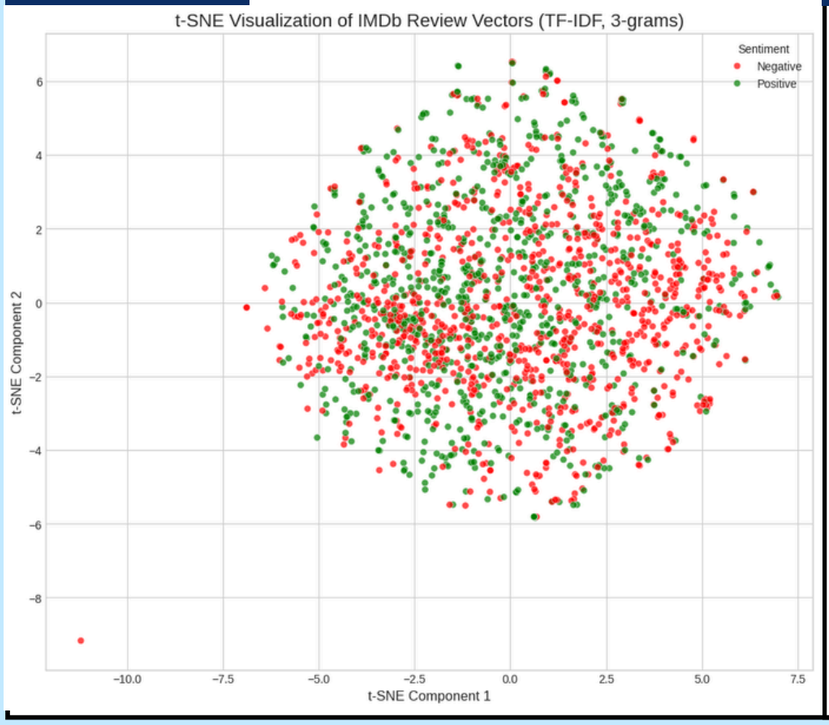
Project Overview

- **The core objective was to build a highly accurate Machine Learning (ML) classifier to automatically determine the sentiment of IMDb movie reviews. The challenge lies in classifying unstructured text (reviews) into a binary output: Positive (1) or Negative (0), overcoming the ambiguity and complexity inherent in human language.**
- **Dataset:** IMDb Movie Review Dataset (50,000 reviews).
- **Preprocessing:** Cleaning (removing HTML, punctuation), lowercasing, and removal of common stop words (NLTK).
- **Feature Extraction:** Text was vectorized using TF-IDF (Term Frequency-Inverse Document Frequency) with a focus on 3-grams to capture local context and phrases (e.g., "not good").
- **Conclusion:** The SENTI-MD pipeline, anchored by the Logistic Regression model and TF-IDF (3-gram) features, achieved a high classification accuracy of **88.4%**. This result validates the chosen NLP approach as highly effective for large-scale, automated sentiment classification of movie reviews.

Illustrations



t-SNE plot



Graphs